DOCUMENT RESUME

ED 172 177 CS 004 886

AUTHOR Bourque, Mary Lyn
TITLE Specification and Validation of Reading Skill and Hierarchies.
PUB DATE Apr 79
NOTE 67p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Classification; Criterion Referenced Tests; Elementary Education; *Organization; Phonics; *Reading Research; *Reading Skills; *Reading Tests; *Research Methodology; Structural Analysis

ABSTRACT

A study was conducted to compare two methodologies for establishing hierarchical relationships--the Dayton and Macready model and the White and Clark procedure--using criterion referenced test data. Several phonics and structural analysis skills were selected from a reading skills battery. The opinions of reading "experts" were solicited in order to establish the hypothesized hierarchy. Both methodologies were then applied to validate the connections. The hypothesized hierarchies were then compared to those produced using empirical data based on the administration of four criterion referenced test levels to approximately 14,000 elementary school children. Specification of the hierarchies based on empirical data via either probabilistic model produced more stable results than those hypothesized by the content specialists. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED172177

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Specification and Validation of Reading Skills and Hierarchies [1, 2, 3]

Mary Lyn Bourque [4]

Providence Public Schools

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Mary Lyn Bourque

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS004886

Running Head: Reading Hierarchies

Abstract

The purpose of this study was to compare two methodologies for establishing hierarchical relationships *viz*, the Dayton and Macready model (1976), and the White and Clark procedure (1973), using criterion referenced test data. Several phonics and structural analysis skills were selected from a reading skills battery. The opinions of "experts" in the field of reading were solicited in order to establish the hypothesized hierarchy. Both methodologies were then applied in order to validate the connections. The hypothesized hierarchies were then compared to those produced using empirical data based on the administration of four criterion referenced test levels to approximately 14,000 elementary school children in an urban setting. Specification of the hierarchies based on empirical data via either probabilistic model produced more stable results than those hypothesized by content specialists.

Author Vita

Mary Lyn Bourque, Research and Evaluation Planner for the Providence Public Schools, received her doctorate in psychometric and evaluative research from the University of Massachusetts, Amherst, in 1979. Dr. Bourque taught for thirteen years in the Boston area, and in recent years has served as a consultant to several educational agencies in New England. She is the author of numerous technical reports in the area of testing and evaluation.

Specification and Validation of Reading Skills and Hierarchies

There is growing concern among practitioners and academicians alike for the numbers of elementary and secondary school pupils who are unable to read and compute at the termination of their school experiences. This concern has sparked a growing interest in instructional and measurement systems research. One topic which is of particular interest is that of learning hierarchies in reading and language arts. Learning hierarchies have been the structure of basal programs as well as reading management systems for a number of years. However, one of the weak aspects of both is the lack of research which focuses on the specification and validation of such hierarchies.

Because the issue of sequencing reading skills is a critical one having serious implications for both the theoretician and the practitioner, this study has focused on the feasibility of specifying and validating such hierarchies using several alternate methodologies.

Specifically, the purpose of this study was three-fold:

1. To establish, using empirical test data, whether or not a hierarchy exists among selected reading skills;
2. To establish the direction of the hierarchy and the strength

of the relationship among the component reading skills;

3. To compare several empirical methodologies for establishing hierarchical relationships.

Hierarchies are usually established and validated in one of three ways or combinations thereof. First, an a priori ordering can be generated. In the case of reading, this type of ordering is usually arrived at deductively based on reading theory. The basal reader and reading achievement tests are probably the best examples of this type of ordering. In many instances the ordering is subsequently verified by the reading practitioner.

A second approach to generating reading hierarchies utilizes psychometric data and is classified as empirical. This type of ordering is usually based on test data and is confirmatory in nature. That is, a specified a priori ordering is validated using appropriately generated and sampled item-sets. The data are expected to be confirmatory of the hierarchy.

Finally a third approach can be classified as experimental. That is, given an existing (or hypothesized) a priori ordering, an instructional unit is designed to teach to the hierarchy. Students are then tested on a pre/post instructional basis for skill acquisition in the hierarchy; hierarchical relationships are then established.

This study was limited to the empirical mode of establishing and validating a reading hierarchy among selected, low-level decoding skills. When a hierarchy was found to exist, the study pursued a more detailed analysis of the hierarchy. That is, an attempt was made to establish

the direction and strength of the relationship among the component skills suggested by the data. Finally, various psychometric methodologies for validating learning hierarchies were compared and the advantages of each approach where possible were pointed out.

Theoretical Background

There have been many significant contributions to the literature on hierarchy specification and validation over the past two decades. This discussion will briefly review the major published studies, focusing on the techniques employed, the inherent weaknesses in the methodologies if any, and the primary results which catalyzed further research. Because the similarities and differences among the methods can be subtle, Table 1 displays a brief review of each method discussed.

Insert Table 1 about here

Gagné and Co-workers

Gagné and his co-workers (Gagné, 1962, 1968; Gagné, Mayor, Garstens & Paradise, 1962; Gagné & Paradise, 1961) addressed the problem of hierarchy validation with the primary question, "What must an individual know or be able to do in order to achieve successful performance on this *class of tasks*, assuming that he is given instructions indicating the form of the desired response and stimulus definitions?" (p.4). Gagné's model is based on the theory that learners cannot successfully perform a task higher in the hierarchy until they have mastered each of the

prerequisite lower tasks. Therefore, beginning with the hypothesized highest task in the hierarchy, the primary question is asked, and then reiterated for each successive lower task. This process is continued until a point is reached where one can be relatively certain of basic skill acquisition by the population. Once the hypothesized hierarchy has been structured, an instructional unit is prepared to teach the skills. Upon completion of instruction, examinees can be divided into four categories:

1. Those who pass the higher skill and all lower skills (11)
2. Those who fail the lower skill and all higher skills (00)
3. Those who fail the higher skill and pass at least one of the lower skills (10)
4. Those who pass the higher skill and fail at least one of the lower skills (01)

Ideally, the 01 category should be the null set. That is, if positive transfer of learning is operating then there should be no examinees who would pass a higher order skill but fail a lower order one. The index "proportion of positive transfer" (PPT) is the ratio of cells (11) + (00) to cells (11) + (00) + (01). Quite obviously the ratio approaches unity as cell (01) approaches zero. Gagné estimated that the values of the PPT index at chance level would be between .25 and .50 (Gagné & Paradise, 1961, p. 9). However, White (1973, p 363) has demonstrated that completely independent skills in a hierarchy whose correlation is zero can achieve an index value of .67, clearly above chance level by Gagné's definition.

White (1973) has also pointed out other weaknesses in this model. First, only one item per skill was used in the study. Consequently, it was not possible to estimate the errors of measurement associated with the exceptions to the hierarchy. Therefore, one is left in a state of indecision relative to establishing whether or not the observed exceptions reflect true deviations in the hierarchy or chance error in the measurement of true connections.

Secondly, the measurement of each skill was delayed until completion of the instructional unit. This allowed for forgetting effects to impact on the results. Pupils could have learned both lower and higher skills and forgotten the lower ones before testing occurred. This would increase the frequency count in the (01) cell thus yielding lower indices which could invalidate true connections in the hierarchy.

Thirdly, Gagné's model can lead to invalid connections in the hypothesized hierarchy. This occurs because the nature of the model is one that is confirmatory. That is, the PPT index, which may be spuriously high, can only confirm connections which are postulated; it cannot suggest connections which may have been overlooked.

Finally, since the sampling distribution associated with Gagné's index is unknown, the variability of the index from one sample to another is also unknown; therefore, the usefulness of the index is limited.

Resnick and Wang

Several years later Resnick and Wang (1969) applied the technique of Guttman's multiple scalogram analysis (MSA) to the validation of learning hierarchies. Their approach was fraught with as many insoluble problems as was

Gagné's. Guttman scales, devised for quite dissimilar purposes, enjoy one unique property: the "parallelogram pattern" of responses, which is both the necessary and sufficient condition for a perfect scale. "The existence of such a 'perfect' scale, or an acceptable approximation to it, is taken to confirm the existence of a behavior hierarchy" (Resnick & Wang, 1969, p. 17). The statistical test employed was the conventional coefficient of reproducibility,

$$\text{Rep} = 1 - \frac{(01)}{(11) + (00) + (01) + (10)} \qquad (1)$$

where the fraction is simply the ratio of the number of errors to the total number of responses. Conceptually, the coefficient of reproducibility is really a measure of the degree to which the scale in question approaches the perfect Guttman scale. The Guttman coefficient used here is quite different from Gagné's statistic insofar as the coefficient of reproducibility applies to the hierarchy as a whole and not to individual connections as does Gagné's index. Resnick and Wang used the Lingoes (1963) modification of scalogram analysis which allows for the generation of multiple scales and thus, branching hierarchies.

However, while in some respects the Resnick and Wang approach may represent an improvement in the technology of hierarchy validation, their research has been criticized by more recent contributors to the field (cf. Airasian & Bart, 1975; White, 1973). First, the Resnick and Wang study tested kindergarten students on skill acquisition without having provided any prior instructional exposure. With this approach

the test is merely a measure of pre-school achievement, and in no way does it validate or reject a given hierarchy. Consequently, the sequence of skill acquisition is unknown and the possibility of random forgetting is largely ignored. Additionally, a very small sample size was used. The general lack, therefore, of these experimental controls leads only to a state of uncertainty about the hierarchy. Moreover, although MSA can generate multiple scales, they are restricted to being independent and linear. That is, once the analysis has indentified a skill as "scaleable" with respect to other higher and/or lower skills, that selected skill may not appear in any other place. This, of course, places severe limitations on the technique and its ability to handle branching within a hierarchy. To overcome this limitation it becomes necessary to test every possible linear scale separately--a cumbersome and costly process.

Finally, as in the Gagné model, there is no sampling distribution because the measurement is limited to one item per skill. Therefore, there is no way of estimating errors of measurement and of generalizing beyond the sample.

Airasian

The model suggested by Airasian (1971) is based on the frequency with which examinees attain a true score (expected) pattern of responses. That is, in a 4-item test with 16 total possible response patterns, only n+1 patterns can be assigned a probability of occurrence greater than zero. Those patterns are 0000, 1000, 1100, 1110 and 1111. For each of the four items, therefore, a ratio of the number of examinees

who attain true score (expected) patterns is calculated. The conditional item difficulty index derived for each true score pattern can be stated generally as

$$P_{i+1} \mid i = \frac{p_{i+1}}{p_i} \qquad (2)$$

where p_i = proportion of examinees who attain the expected or true score response patterns.

Airasian admits to at least two serious confounding influences on his model: (a) item validity is critical, and (b) the index is highly sensitive to the sequencing of instruction. Item validity is a problem with each of the models discussed thus far. In addition, the Airasian model does not solve the problem discussed earlier relative to estimating errors of measurement; it does not provide for a sampling distribution; and it largely ignores the forgetting and/or guessing parameters.

Walbesser and Eisenberg

The dependency test ratios of Walbesser and Eisenberg (1972) are not unlike Gagné's proportion of positive transfer. Walbesser proposes five ratios for specifying hierarchical relationships:

1. Consistency ratio: ratio of examinees mastering the higher skill to those who have mastered the lower skill: $\frac{(11)}{(11) + (10)}$ (3)
2. Adequacy ratio: ratio of examinees mastering the lower skill to those who have mastered the higher skill: $\frac{(11)}{(01) + (11)}$ (4)

3. Necessity ratio: ratio of examinees who failed the lower skill and also failed the higher skill:

$$\frac{(00)}{(00) + (10)} \quad (5)$$

4. Completeness ratio: ratio of examinees who mastered the higher skill to those who both mastered or failed both skills completely:

$$\frac{(11)}{(11) + (00)} \quad (6)$$

5. Inverse Consistency ratio: ratio of examinees who failed the lower skill to those who mastered the higher skill:

$$\frac{(00)}{(00) + (01)} \quad (7)$$

Walbesser has arbitrarily set a lower limit of .85 for the consistency, adequacy and completeness ratios. If these ratios do not exceed this threshold value then a valid hierarchical connection cannot exist.

Both Capie and Jones (1971) as well as White (1974c) have been critical of the Walbesser model. First, Capie and Jones point out that errors of measurement are likely to suppress both the consistency and necessity ratios; and that the latter is quite sensitive to the number of failures on the test. Consequently, the ratios are also sensitive to low reliability and validity indices. White (1974c) also points out that the ratios lack generalizability; and that high ratios suggest invalid hierarchical relationships. White's final conclusion is that the ratios are more nearly a measure of the degree of difficulty of the content for the examinees. "Their uselessness of hierarchy validation is illustrated by the fact that only five of the 31 connections in three

of Gagné's studies [1961, 1962, 1965] meet the condition that all three indexes should be 0.85 or greater, and four of these five involve lower skills which were achieved by all Ss" (p. 63).

Capie and Jones

The work of Capie and Jones (1971) was primarily a comparative study of the dependency test ratios of Walbesser and Eisenberg (1972) with the phi-correlation coefficient. They based their argument on the fact that the skills in a true hierarchy should clearly demonstrate increasing correlational patterns if one were to compute product moment correlations between individual skills and the total number of skills mastered. This concept is at least intuitively appealing. They also suggest that a parallel set of correlations be derived for a transfer test, based on the premise that increased mastery of skills in the hierarchy must necessarily result in increased transfer of learning. Consequently, increasing correlational patterns should be in evidence for both the transfer test as well as the individual skills if it is a true hierarchy. To compensate for the unreliability of the tests, phi can be corrected for attenuation.

Capie and Jones detail several cautions about their model. First, the technique is highly contingent upon the validity and reliability of the test. If test validity is questionable, or if the tests are so short that reliability is equally dubious, then the technique is severely weakened. More importantly, however, the testing sequence is based on the arbitrary judgment of the researcher. To the extent that such judgment is on target, then the testing sequence will reflect all the

possible prerequisite skills. This is not always the case, however.

Bart, Airasian and Krus

There have been several contributors to the ordering-theoretic approach to hierarchy validation, including Airasian and Bart (1973, 1975), Bart and Airasian (1974), Bart and Krus (1973), and more recently Baker and Hubert (1977).

Bart and Krus (1973) present a variation on a theme detailed by Gagné (1970a). First, let us assume that all items are dichotomously scored: "0" for the incorrect response, "1" for the correct response. We can then define a "prerequisite" relationship as follows: ". . . success on item i is a prerequisite to success on item j if and only if the response pattern (0) for items i and j respectively does not occur" (Bart and Krus, 1973, p. 292). Cells (00), (10), and (11) are defined as *confirmatory* response patterns; while the (01) cell is defined as *disconfirmatory*. For any given set of n items it is possible to construct an nxn matrix where entries represent the percentage of *disconfirmatory* response patterns for each of the two-item sets. It is then necessary to set a tolerance level (Σ) for the percentage of disconfirmatory responses above which a relationship between the item-pairs will be rejected. So, for example, if the researcher has a pre-established tolerance level of 5% and items i *and* j have a 4% disconfirmatory response pattern, then the conclusion is simply that item i is a prerequisite to item j. Naturally, low tolerance levels are recommended so that spurious relationships will not be validated.

More recently Baker and Hubert (1977) have elaborated on the ordering-theoretic model introducing a *directed-graph* representation of the data, and suggesting a statistical procedure for testing the goodness-of-fit of the data to an hypothesized hierarchy which compares all n! possible permutations of the empirically generated prerequisite matrix.

There are at least two major limitations to the ordering-theoretic model. First, it is quite obviously limited to pairwise analysis. Consequently the model more easily accommodates itself to linear rather than branching hierarchies. Secondly, the relationships established are functions of Σ, the pre-established tolerance limit. This is the way in which the model accounts for errors of measurement.

White

White has been one of the sharpest critics of the various models for validating learning hierarchies (cf. White, 1973, 1974a, 1974b, 1974c; White & Clark, 1973). White (1974a) lists five weaknesses inherent in the previously discussed models which limit their usefulness;

1. Loosely or ill-defined instructional objectives;
2. Single item estimates of mastery which preclude estimates of error of measurement;
3. Improper indices for estimating hierarchical relationships;
4. Improper timing of the testing which encourage forgetting or guessing;
5. Small sample sizes employed in several studies.

In order to correct for these five weaknesses, White suggests nine steps that should be implemented in validating a hierarchy.

1. Define the hypothesized highest-order skill in the hierarchy in behavioral terms.
2. Define the remaining lower-order skills by asking Gagné's question vis-à-vis each lower-order skill in turn.
3. Content-validate the hypothesized hierarchy by soliciting the opinion of "experts".
4. Postulate sub-divisions of each skill in the hierarchy so that skill definitions obtain a high degree of precision.
5. Test empirically whether or not these new subdivisions really do represent unique skills.
6. Develop an instructional unit for teaching skill acquisition, embedding mastery tests in it for each skill.
7. Use a minimum of 150 examinees allowing them to work through the instructional unit and to be tested as appropriate.
8. Using frequency counts in a 2x2 matrix, examine the results for large numbers in the (01) cell which would indicate rejection of the postulated connection.
9. Modify the hierarchy to include only those connections that survived the validation process.

White's model is a carefully thought-out process that addresses the major weaknesses found in other models, viz., ill-defined elements, errors of measurement, statistical tests of goodness-of-fit, forgetting and guessing parameters, and small numbers of examinees. On the other hand, an experimental model such as this is quite difficult to implement within the constraints the researcher usually finds in conducting a

field-study.

One aspect of the model requiring further elaboration is that of the test of inclusion (White & Clark, 1973). This is the only model that can accommodate multi-item data-sets for each objective in the hierarchy and, consequently, can estimate the probability of a randomly selected examinee having answered zero, or one or more times correctly for any objective-pair. White and Clark develop an example for the two-item and three-item case.

Assume a matrix of response frequencies as indicated in Figure 1.

Insert Figure 1 about here

White and Clark define the population sub-groups as follows:

P_0 = proportion of the population having neither skill

P_B = proportion of the population having both skills

P_I = proportion of population having only Skill I

P_{II} = proportion of population having only Skill II

Likewise, the conditional probabilities of members of the group answering the appropriate number of items correctly are:

θ_a = probability of examinee with Skill I answering correctly any item for Skill I

θ_b = probability of examinee without Skill I answering correctly any item for Skill I

θ_c = probability of examinee with Skill II answering correctly any item for Skill I

θ_d = probability of examinee without Skill II answering correctly any item for Skill II

For each cell in the matrix the total probability is defined as the product of the probability of the examinee's being in each group and the conditional probability of members of the group answering the appropriate number of items correctly. A probability estimate for each cell in Figure 1 reflecting the nine possible outcomes could be derived if the values of the various Ps and θs were known. For each cell in the matrix estimates of P, the proportion of the population who have neither skill, both skills, or one or the other skill, as well as θ, the conditional probabilities, can be derived either from the cell frequency using a maximum likelihood procedure or from the marginal totals. In the two-item case P_{02} can be found by substituting the estimates of P and θ in

$$P_{02} = P_0(1-\theta_b)^2\theta_d^2 + P_I(1-\theta_a)^2\theta_d^2 + P_{II}(1-\theta_b)^2\theta_c^2 + P_B(1-\theta_a)^2\theta_c^2 \quad (8)$$

Estimates could be calculated for the remaining eight cells, allowing for the generation of a multinomial distribution, and the test of significance could look for deviations outside the observed distribution. However, all this is not necessary according to White and Clark. They suggest that the P_{02} cell would sustain the greatest proportional change if shifts in frequency obtained in the other eight cells. Therefore, it is merely necessary to set a critical value, C, beyond which H_0 will be rejected if the observed frequency, f_{02}, in the (02) cell is in excess of C. While White and Clark do not suggest real values for C they do point out that the magnitude of C is directly pro-

portional to a large N and a correspondingly high value of P_{II}. The reasoning and mathematical development is parallel for the three-item case.

Dayton and Macready

Dayton and Macready have developed three models for hierarchy specification and validation (Dayton and Macready, 1976a, 1976b; Macready, 1975; Macready and Dayton, 1977; Macready and Merwin, 1973). Model I is a classification (mastery/non-mastery) model with error probabilities associated with each individual item. Model II is also a classification model wherein error probabilities are constant across items.

Model III which was utilized in this study is a generalized version with constant guessing and forgetting parameters wherein $\mathbf{u}$ is a column vector representing the examinee's response pattern, comprised of 0's and 1's; and $\mathbf{v}$ is a pattern vector which, as a set, represents the response pattern for the a priori hierarchy, comprised of 0's and 2's. The probabilistic model may be written as

$$P(\mathbf{u}) = \sum_{j=1}^{q} P(\mathbf{u}|\mathbf{v}_j) \cdot \Theta_j \quad \text{and} \tag{9}$$

$$P(\mathbf{u}|\mathbf{v}_j) = \prod_{i=1}^{n} \alpha_i^{a_{ij}} (1-\alpha_i)^{b_{ij}} \beta_i^{c_{ij}} (1-\beta_i)^{d_{ij}} \tag{10}$$

where α and β are guessing and forgetting parameters.

Let g_{ij} = vector difference, $\mathbf{v}_j - \mathbf{u}$. Then

$a_{ij} = 1$ if $g_{ij} = -1$; 0 otherwise. (Examinee guessed)

a_{ij} = 1 if g_{ij} = 0; 0 otherwise. (Examinee didn't know item and responded incorrectly)

c_{ij} = 1 if g_{ij} = 2; 0 otherwise. (Examinee forgot)

d_{ij} = 1 if g_{ij} = 1; 0 otherwise. (Examinee knew item and responded correctly)

For any given objective, three of the above will equal zero, the fourth will assume the appropriate value: 1, -1, 0 or 2. The model allows for the estimation of 2^n-1 independent parameters where n equals the number of items. However, only 2^n-2 may be fitted to any model in order to provide for the goodness-of-fit test. Once the parameters have been estimated, the expected proportions for each 2^n data vectors can be computed. At this point either the chi-square test or likelihood ratios is appropriate for estimating goodness-of-fit.

A computer program, known as MODEL 5, was developed by the authors for computing all parameter estimates and testing significance. The number of items is limited to less than ten, but more than three with no more than one item per task. In general, both the Dayton and Macready Model as well as the White and Clark Model represent a significant advance in the technology of hierarchy specifications and validation. Both of the models, because they are probabilistic rather than deterministic, allow for the estimation of errors of measurement, and therefore, for statistical tests of fit. The Dayton and Macready Model also allows for specific errors due to guessing and forgetting. Both models have advantages that are worth noting. Table 2 summarizes the results of such a comparison.

Insert Table 2 about here

Some Summary Comments

This review has briefly presented the critical aspects of eight methodologies employed in the specification and validation of learning hierarchies. Each model varies considerably in its technological sophistication for dealing with the inherent problems of hierarchy validation research. Some of the earlier models for example have considerable intuitive appeal because of their simplicity, while the later models capitalize on recent advances in computer technology. The "state of the art" is far from having reached closure on the issue. Only continued research in the area will produce the desired results, viz., a straightforward, perhaps eclectic, model which minimized the weaknesses and maximizes the strengths of each of the preceding methodologies.

Methodology

Design

Table 3 specifies the 13 phonics and structural analysis skills finally selected for this study and measured by *The Reading Skills Inventory: A Criterion Referenced Assessment* (Hambleton, 1975).

Insert Table 3 about here

The phonics and structural analysis sub-hierarchies were considered separately as well as by clusters within each sub-hierarchy. Clustering was necessary because the selected objectives were measured across four different test levels while examinees were only given any two contiguous levels.

Sample of Examinees

The *Inventory* was administered to approximately 14,500 examinees in grades one through six in one small urban district in the Northeast. Examinees were primarily from families of low to middle socioeconomic level in a blue-collar community. Additionally, the community is strongly homogeneous in ethnic composition, though desegregated having approximately 40% total minorities.

The assignment of test level for each examinee was made on the basis of two criteria: (a) the pre-test level which the examinee had taken 12 months earlier, and (b) teacher-judgment. Where possible, the test level assigned to an examinee was that level which the student is known not to have mastered. This judgment was based in large measure on the pre-test data available on each examinee. Where pre-test data were not available, as in the case of an inter-district/intra-district transfer or an absentee, or when, in the best judgment of the pupil's teacher, another level was clearly more appropriate, the test level assignment depended primarily on teacher-judgment.

Instrumentation

The Reading Skills Inventory: A Criterion-Referenced Assessment was designed to assist the implementation of a program of individually-paced reading instruction in grades K through 8. One of the two components of the *Inventory* consists of a set of criterion-referenced word-attack measures at each of five instructional levels: pre-reading, Level I, II, III, and IV. The original test specifications were generated based on the district's reading curriculum scope and sequence.

The decision regarding the inclusion or exclusion of a particular objective was made primarily by judgment sampling based on the professional opinion of reading specialists and teachers in the district. The primary purpose to which the test results would be directed was the tracking of individual student performance. Secondarily test data would also determine ad hoc classroom groupings, as well as influence decisions relative to student pacing.

A validation study conducted by the author of the *Inventory* (Hambleton, 1975) established the adequacy of each level in the battery. Validation data included both an analysis of item difficulties and item discrimination indices. Reliability assessment included measures of internal consistency; parallel form reliability; and proportion of agreement of mastery decisions based on two performance standards (80% and 100%). In addition a content validation study was conducted.

Procedures

The *Inventory* was administered over a three week period in the Spring. Most items on the test could be group-administered. However, some objectives required a one-to-one administration. These were given by the regular classroom teacher as time allowed during the testing period. Test results were then sent to a service bureau for hand or machine scoring. All items were scored dichotomously: "0" for a fail, and "1" for a pass. For each objective, a cutting score of either n-1 or n-2 was used. This was particularly appropriate in implementing the Dayton and Macready model. For the White and Clark model the item scores were selected on a random basis.

In order to establish an a priori hierarchy based on the judgment of experts in the field of reading a sample of 23 content specialists was asked to respond to a pair-wise comparison task. Each respondent examined 56 pairs of objectives resulting from two 8-objective clusters: one cluster of phonics skills, and a second of structural analysis skills.

The respondent was asked to evaluate for each objective-pair whether mastery of the first objective of the pair was necessary prior to the mastery of the second objective; and if so, which objective should be mastered first. If the respondent felt that the relationship between the objectives was not a necessary one but merely a facilitating one, then the respondent indicated which objective would facilitate the other. Finally, the respondents were allowed the option of indicating no relationship between the objectives; that is, each objective could be mastered independently of the other.

In order to insure that all respondents clearly understood the taks, two bogus objectives were included in the sequence, thus increasing the number of pairs evaluated to 72. The two respondents eliminated from the sample failed to correctly classify the bogus objectives in the third option, i.e., these objectives could clearly be mastered independently of all other objectives in the cluster. The hierarchies generated through expert judgment were then compared to those produced using empirical data based on the administration of the four criterion-referenced test levels. Initially the Dayton and Macready model for specifying a hierarchy utilizing a maximum likelihood solution was applied.

Secondly, the White and Clark procedure, a pair-wise comparison method having a "test of inclusion" significance test, was applied. This procedure can accommodate multi-item data-sets for each objective in the hierarchy, and, as a result, can estimate the probability of a randomly selected examinee having answered zero, or one or more items correctly for any objective-pair.

Computer Programs

George M. Macready of the University of Maryland provided a listing of the Dayton and Macready program, MODEL 5, which was used to conduct that segment of the data analysis. No substantive changes were made in the program, only those necessary for compatability with the CDC system.

It was necessary to write a program, INCLU, based on the White and Clark model. The program allows for either the two-item or three-item pair-wise comparison, with a sub-routine for randomly selecting items for each comparison, and accepts scored item data for the input data set.

Results

Expert Judgment Hierarchies

One of the initial efforts of this study was to establish an a priori ordering of the objectives based on the collective judgment of professionals in the field of reading. Given the wide variability in response patterns, and given the possibility that the hierarchy specification task could have been perceived differently by each of the 21 respondents, it seemed advantageous to collapse objective-pair data across response patterns. This procedure, while possibly establishing the direction of the relationship, leaves the strength of the relationship undecided.

Table 4 summarizes the collapsed results of all objective-pair judged by the experts.

Insert Table 4 about here

Clearly, patterns emerge and it appears possible to suggest tentative hierarchies among the objective-pair for each data set. Figure 2 displays the possible hierarchical relationships generated by expert judgment. The reader will note that for Data Set I three options are provided, while for Data Sets IV and V only one hierarchical arrangement is suggested.

Insert Figure 2 about here

Using the response patterns from Table 4 it was possible to calculate the percent of agreement between the judgment data and the proposed hierarchies of Figure 2. The index of agreement was derived by estimating the percent of responses favoring the objective's position in the hierarchy based on the total number of judgments for that hierarchy. In Table 5 the reader will note that the indices range from a low of 25% to a high of 64%. However, the average for a given hierarchy is usually greater than 50%. In the case of Data Set I, two of the proposed hierarchies have the same average index of agreement; one is somewhat lower. Data Set V has the lowest average index, 44.4%. The depression of this average index is due to the inclusion of objective 64 in the

hierarchy, the position of which enjoys only a 25% agreement with the judgment responses.

Dayton and Macready Hierarchies

The first of two empirical procedures utilized to validate the hierarchies was that of Dayton and Macready. The criterion score which determined mastery was n-1 on the first pass and n-2 on the second, where n equalled the total number of items measuring a given objective. Table 6 summarizes the maximum likelihood estimates of the parameters and their standard errors at each criterion level for each data set.

Insert Table 6 about here

The frequency data from these parameter estimates were generated for all possible response patterns for each data set. Space limitations obviously preclude displaying in table form all values obtained. However, for purposes of clarity Table 7 summarizes the observed and predicted frequencies, chi-square and probability estimates for the true-score patterns for each data set at both criterion levels.

Insert Table 7 about here

In all cases the values for the guessing and forgetting parameter were unrestricted. Using the MODEL 5 program, a solution was obtained (convergence to a criterion of .1E-04) for Data Sets I and IV on both passes. For Data Set V the maximum number of iterations (100) was

reached without attaining convergence. This solution, therefore, is questionable. The user-specified true score patterns reflected a linear hierarchy in all cases.

Table 8 summarizes the chi-square goodness of fit statistics for each data set at both criterion levels. The significance of the Pearson Chi-Square is applicable, of course, only to the complete data set. And as the reader will note, all are significant. However, inspection of the discrete chi-square variable in Table 7 shows several true score patterns having non-significant outcomes. In Data Set I, for example, at the n-1 criterion score all but the last true score pattern (111111) have non-significant chi-square values; at the n-2 criterion score only the last two (111110 and 111111) are significant while the remaining five are not. In Data Set IV only the first pattern (0000) at the n-2 criterion score is non-significant; all others are. In Data Set V three patterns (10000, 11100, and 11111) at the n-1 criterion score are non-significant; all remaining patterns are. Since the distribution chi-squares are significant while some of the discrete chi-squares are not, the implication seems to be that for the entire data set the sample size is so large that it would be impossible not to achieve significance. In other words, a large N has led to the rejection of the null hypothesis. Notwithstanding the fact that the compatability between the observed and the theoretical frequency distribution is low, a consideration of the value of the discrete chi-square variable for each true score pattern leads to the specification of tentative hierarchies for some of the data sets. These are displayed in Figure 3.

Insert Figure 3 about here

White and Clark Hierarchies

The second empirical procedure utilized to validate the hierarchies was that of White and Clark. This procedure is based on the pair-wise comparisons of all objectives in any given sequence utilizing either two or three items per objective to assess mastery. The program INCLU was used to generate the frequencies based on a random selection of two items per objective on the first pass and three items per objective on the second. In all other respects the data sets remained the same except the way in which the response patterns were analyzed. For the two-item case, three replications were generated, while for the three-item case, two replications were generated. In the interest of space, it is not possible to reproduce all the statistics for each of the data sets. However, INCLU generated the following statistics: the estimated probability of the "02 event," i.e., the probability that a member of the sample will be classified in the I-0, II-2 cell; the mean and standard deviation for the given distribution; a critical value "C," if when exceeded by the observed frequency in the 02 cell, will cause H_o to be rejected: and the observed frequency, f_{02}, i.e., the actual number of cases classified in the I-0, II-2 cell. A parallel set of statistics was generated for the "03 event".

In examining the $f_{02/03}$ values for each objective-pair and suggesting tentative hierarchical relationships, the following decision

rules applied:

1. For any one objective-pair, H_o was rejected if $f_{02/03} > C$; H_o was accepted if $f_{02/03} \leq C$.
2. Over five replications, if H_o was rejected 80% or more then no relationship in the objective-pair was acknowledged.
3. Objective-pairs were rank-ordered according to their percent of rejection, ranging from 0% to 60%.
4. Lower rejection-rate pairs were established in the hierarchy first followed by higher ones until all were accounted for.

The resulting hierarchies are displayed in Figure 4.

Insert Figure 4 about here

In Data Set I, two objectives (24 and 25) were rejected from hierarchical consideration by decision rule one on two occasions. This resulted in two possible structures. In Data Set IV, again two objectives (55 and 56) while not rejected by decision rule one retained their position within the hierarchy with only weak supporting evidence. Therefore, it was deemed advisable to suggest two alternate structures. Data Set V seemed to present clear evidence in favor of the structure provided in Figure 4.

Discussion

Data Set I: Phonics Hierarchy

Data Set I is the largest of the three examined in this study, and thus, presented its own problems simply because of the total number of objectives in the sequence. The reader will note in Figure 5 that the hierarchies displayed vary in the total number of objectives included.

Insert Figure 5 about here

Those derived from "expert judgment" contain all six phonics objectives, with three variations on a theme. Those derived through the Dayton and Macready procedures omit ending consonant digraphs and/or beginning consonant digraphs. Those resulting from the application of the White and Clark model omit ending consonant sounds in one instance and auditory discrimination (rhyming) in the second case. There are some areas of agreement among the three. It appears that digraphs (beginning or ending or both) are at the bottom. What comes in between is debatable, with vowel sounds and auditory discrimination (rhyming) as the potential candidates.

It seems also clear from the data that beginning sounds (either digraphs or consonants) precede ending sounds. The position of vowel sounds in the hierarchy is likewise worth noting. The position of this particular objective was clearly undecided by the experts and, if placed anywhere in the hierarchy, it was generally thought to be a branch of

the primary hierarchy. This is quite contrary to the evidence provided by the two empirical models. The position of the auditory discrimination objective is somewhat surprising. One might assume that, as a listening skill, it falls into the pre-reading or readiness category, and as such, would have the lowest position in the hierarchy of formal reading skills. However, this does not appear to be the case. In both the judgment of the experts, as well as by the empirical evidence, this skill may be, and indeed is, preceded by beginning consonant sounds.

Data Set IV: Structural Analysis Hierarchy

Data Set IV has four objectives in total representing the structural analysis domain. Figure 6 presents the proposed hierarchical structures.

Insert Figure 6 about here

The reader will notice that the expert judgment hierarchy omits the syllabication objective. Since the Dayton and Macready procedure failed to identify any relationships in this data set that were significant, this comparison will confine itself to the White and Clark model only. Clearly there is little evidence to support agreement between the a priori hierarchy and either of those resulting from the White and Clark procedures. Only the relationship betwen root word derivations and root words + affix seems clear, with the former lower in the hierarchy than the latter in both models.

Generally the failure of the Dayton and Macready procedure to identify structures, coupled with the lack of agreement between White

and Clark and expert judgment, would lead the author to suggest that no true hierarchy exists among this sequence of structural analysis skills.

Data Set V: Structural Analysis Hierarchy

This data set consisted of five objectives also from the structural analysis domain. Figure 7 displays the resulting hierarchies suggested by the three procedures.

Insert Figure 7 about here

The expert judgment hierarchy omits one objective from the sequence, viz., syllabication of nonsense words. Two are ommited from the Dayton and Macready sequence, viz., suffixes and syntax (verbs) and prefixes/suffixes. Only the White and Clark sequence maintains all five objectives. Notwithstanding the omissions, the degree of agreement among the three hierarchies is remarkably high. Two of the three agree that prefixes/suffixes is the lowest level skill of the five under consideration. All three agree that root word + affix is also a low level skill. Syllabication of nonsense words assumes the mid-position in two of the three hierarchies. Suffixes and syntax (nouns, verbs or both) top the list for all three models.

Conclusions

First, let us examine the hierarchy specification task performed by content experts. The basic question is, do experts in the field of reading admit of a hierarchy existing among a selected number of low-level phonics and structural analysis skills. Generally, the data seem to indicate that the experts concur on only one aspect of hierarchical

relationships, viz., sequence. That is, the experts indicate a willingness to say Skill I precedes Skill II. However, they stop short of saying that Skill I *must* be mastered before Skill II *can* be mastered; or that the mastery of Skill II will be facilitated by the prior mastery of Skill I; or the mastery of both skills is mutually dependent, one on the other. In the final analysis, the hierarchies resulting from the judgment of experts were at best weak. Although the author has suggested some hierarchical structures, they should be viewed cautiously.

The hypothesis of interest in the Dayton and Macready procedure was

$$H_o: \; P_{j_{obs}} = P_{j_{pred}}$$

$$H_1: \; P_{j_{obs}} \neq P_{j_{pred}}.$$

Clearly, the null hypothesis must be rejected based on the chi-square results given in Table 8. None of the data sets lends itself to a hierarchical interpretation based on a linear sequence of true score patterns. Examination of the values of α and β shows that only for Data Set I are these parameters within expected bounds. The unusually high value of the "forgetting" parameter in Data Sets IV and V can possibly be interpreted in several ways. First, and the most obvious, is that a linear sequence just does not provide a reasonable fit for the data. A second possible interpretation is that examinees were not appropriately assigned to consecutive test levels. While this interpretation is not so obvious, it is, in the opinion of the author, one of the most likely reasons for the apparent lack of fit of the data. Further substantiating evidence is provided for this interpretation by the relatively high

values of the "guessing" parameter as well. Also, it should be noted that as the criterion score for mastery was lowered from n-1 to n-2, the values for both parameters began to approach more reasonable limits. When the criterion score is a composite of two or three items, as in the White and Clark model, the degree of fit is improved markedly. One might conclude, therefore, that the fewer the items the better the fit, which in this case, is quite true. However, this is contrary to what we already know about the assessment of mastery and the required number of items necessary to minimize the number of misclassifications.

In the White and Clark procedure the hypothesis of interest was

$$H_0: \ f_{on} \leq C$$

$$H_1: \ f_{on} > C$$

where f_{on} has been previously defined as the observed frequency in the I-0, II-2 cell in the two-item case and I-0, II-3 in the three-item case; and C is a selected threshold value, which if exceeded, would exclude Skill I and/or II from the hierarchy under consideration. In this particular study the value of C was further defined as the nearest integer point to the third standard deviation above the mean of the sampling distribution. It should be noted that the normal approximation to the binomial distribution was used in order to simplify the estimation of the distributional statistics. Because of the multiple replications it was necessary to establish the five decision rules defined in the preceding section. On that basis it was possible to arrive at the hierarchical structures displayed in Figures 5, 6, and 7. This model is really a special case of the general probabilistic model, as pointed out by

Dayton and Macready (1976a), which utilizes marginal frequencies instead of a maximum likelihood estimation procedure.

The application of the White and Clark model appears to be less sensitive to some of the overall distributional problems found in the Dayton and Macready approach. Consequently, the hierarchical structures emerged without undue strain on the data. Now, of course, at least two assumptions were made by the researcher in addition to those indigenous to the model itself: (1) that each replication was an equally reliable estimate of the probability of the "02/03 event"; and (2) that all items were equally reliable measures of mastery of the consequent skill. Both assumptions seem reasonable in view of the fact that test items were generated on the basis of domain-specifications, and that their order of presentation to the examinee was random and not according to item-difficulty indices. In fact, the resulting structures match fairly well with the a priori hierarchies generated by the content experts.

It is clear from a consideration of the results of this study that White's procedure (1974b) indeed outlines the optimal steps for conducting hierarchy research under properly controlled conditions. Given that those conditions exist it should be possible to compare several alternate hierarchies fitted to the same data. For example, in this study a strictly linear hierarchy was examined. However, it is possible under the Dayton and Macready (1976a, 1976b) model to consider branching hierarchies as well. The focus of any further research in this area should be directed at a rigidly controlled design that precludes the invalidating features of a field-based approach. Within the constraints

of such a design it would be quite possible to examine aspects of item-design and its relation to hierarchy validation.

One might also wish to examine the long-term effects of hierarchical learning. To date, only a few studies have examined the issue of retention of hierarchically learned skills over a period of time (White, 1976 and White and Gagné, 1978). Both of these studies postulate that if learning is hierarchical so also is retention. Whether or not this hypothesis can be applied in the area of reading needs to be explored.

This study has shown that quite possibly learning hierarchies do exist in the area of reading. Future research should focus on limited content domains to affirm or disaffirm this initial evidence. The benefits resulting from the specification of such hierarchies are several. First, validated hierarchical patterns among reading skills should result in improved instructional sequences. If prior acquisition of certain skills is necessary to the posterior acquisition of alternate skills then proper curriculum sequencing becomes critical in the instructional design process. Secondly, validated hierarchies should allow for more efficient and effective diagnosis and prescription. This is particularly important when the practitioner in reading is faced with the problems of remediation. Finally, hierarchical relationships among instructional objectives should foster the development of tailored testing programs as well as improve formative pupil-evaluation procedures.

References

Airasian, P. W. A method for validating sequential instructional hierarchies. *Educational Technology*, 1971, *11*, 54-56.

Airasian, P. W., & Bart, W. M. Ordering theory: A new and useful measurement model. *Educational Technology*, 1973, *13*, 56-60.

Airasian, P. W., & Bart, W. M. Validating a priori instructional hierarchies. *Journal of Educational Measurement*, 1975, *12*, 163-173.

Baker, F. B., & Hubert, L. J. Inference procedures for ordering theory. *Journal of Educational Statistics*, 1977, *2*, 217-233.

Bart, W. M., & Airasian, P. W. Determination of the ordering among seven Piagetian tasks by an ordering-theoretic method. *Journal of Educational Psychology*, 1974, *66*, 277-284.

Bart, W. M., & Krus, D. J. An ordering-theoretic method to determine hierarchies among items. *Educational and Psychological Measurement*, 1973, *33*, 291-300.

Capie, W., & Jones, H. L. An assessment of hierarchy validation techniques, *Journal of Research in Science Teaching*, 1971, *8*, 137-147.

Cotton, J. W., Gallagher, J. P., & Marshall, S. P. The identification and decomposition of hierarchical tasks. *American Educational Research Journal*, 1977, *14*, 189-212.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. *Psychometrika*, 1976, *41*, 189-204. (a)

Dayton, C. M., & Macready, G. B. *Computer programs for probabilistic models*. Unpublished manuscript, 1976. (Available from the authors, University of Maryland, College Park, Maryland.) (b)

Gagné, R. M. The acquisition of knowledge. *Psychological Review*, 1962, *69*, 355-365.

Gagné, R. M. Learning hierarchies. *Educational Psychologist*, 1968, *6*, 1-9.

Gagné, R. M. *The conditions of learning*. (2nd ed.) New York: Holt, Rinehart, and Winston, 1970. (a)

Gagné, R. M. Some new views of learning and instruction. *Phi Delta Kappan*, 1970, *51*, 468-472. (b)

Gagné, R. M., Mayor, J. R., Garstens, H. L., & Paradise, N. E. Factors in acquiring knowledge of a mathematical task. *Psychological Monographs*, 1962, *76*, (7, Whole No. 526).

Gagné, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. *Psychological Monographs*, 1961, *75*, (14, Whole No. 518).

Hambleton, R. K. *The reading skills inventory: A criterion-referenced assessment*. Unpublished pilot edition, 1975.

Hambleton, R. K., & Eignor, D. R. *Adaptive testing applied to hierarchically structured objective-based programs*. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Lingoes, J. C. Multiple scalogram analysis: A set theoretic model for analyzing dichotomous items. *Educational and Psychological Measurement*, 1963, *23*, 3.

Linke, R. D. Replicative studies in hierarchical learning of graphical interpretation skills. *British Journal of Educational Psychology*, 1975, 45, 39-46.

Macready, G. B. The structures of domain hierarchies found within a domain referenced testing system. *Educational and Psychological Measurement*, 1975, 35, 583-598.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 1977, 2, 99-120.

Macready, G. B., & Merwin, J. C. Homogeneity within item forms in domain referenced testing. *Educational and Psychological Measurement*, 1973, 33, 351-360.

Resnick, L. B., & Wang, M. C. Approaches to the validation of learning hierarchies. *Proceedings of the Eighteenth Annual Regional Conference on Testing Problems*. Princeton, New Jersey: Educational Testing Service, 1969.

Resnick, L. B., Wang, M. C., & Kaplan, J. Task analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. *Journal of Applied Behavior Analysis*, 1973, 6, 679-710.

Walbesser, A. H., & Eisenberg, T. A. *A review of research on behavioral objectives and learning hierarchies*. Columbus, Ohio: ERIC Information Analysis Center for Science, Mathematics and Environmental Education, 1972. (ERIC Document Reproduction Service No. ED 059 900.)

White, R. T. Research into learning hierarchies. *Review of Educational Research*, 1973, 43, 361-375.

White, R. T. The validation of a learning hierarchy. *American Educational Research Journal*, 1974, *11*, 121-136. (a)

White, R. T. A model for the validation of learning hierarchies, *Journal of Research in Science Teaching*, 1974, *11*, 1-3. (b)

White, R. T. Indexes used in testing the validity of learning hierarchies. *Journal of Research in Science Teaching*, 1974, *11*, 61-66. (c)

White, R. T. Effects of guidance, sequence and attribute-treatment interactions on learning, retention, and transfer of hierarchically ordered skills. *Instructional Science*, 1976, 5, 133-152.

White, R. T., & Clark, R. M. A test of inclusion which allows for errors of measurement. *Psychometrika*, 1973, 38, 77-86.

White, R. T., & Gagné, R. M. Formative evaluation applied to a learning hierarchy. *Contemporary Educational Psychology*, 1978, 3, 87-94.

Footnotes

[1]This article is based on the author's doctoral dissertation submitted to the Graduate School of the University of Massachusetts in partial fulfillment of the requirements for the doctoral degree. I would like to express my appreciation to Ronald K. Hambleton under whose direction the study was conducted and whose criticisms and suggestions were most helpful in preparing the final manuscript. I am also grateful to George M. Macready of the University of Maryland for providing a listing of MODEL 5 used to analyze the data, as well as Leah Hutton of the University of Massachusetts for generating the program INCLU, and assisting in the data analysis.

[2]A version of this article will be presented at the meeting of the American Educational Research Association, San Francisco, California, April 1979.

[3]*Laboratory of Psychometric and Evaluative Research Report No. 96.* Amherst, MA: School of Education, University of Massachusetts, 1979.

[4]Requests for reprints should be sent to Mary Lyn Bourque, Research and Evaluation Office, Providence Public Schools, 150 Washington Street, Providence, Rhode Island, 02903.

Table 1

An Abridged Review of Hierarchy Validation Methodologies, 1961 - 1977

Author	Statistical Techniques Employed	Weaknesses in the Model
Gagne and Paradise 1961	PROPORTION OF POSITIVE TRANSFER (PPT) $PPT = \frac{(11) + (00)}{(11) + (00) + (01)}$	1. Unable to estimate errors of measurement since there is only one item per learning set. 2. Forgetting parameter not considered. 3. Inflated PPT index leads to invalid connections in the hierarchy. 4. Lacks sampling distribution and thus, generalizability.
Resnick and Wang 1969	MULTIDIMENSIONAL SCALOGRAM ANALYSIS (MSA) $Rep = 1 - \frac{(01)}{(11) + (00) + (01) + (10)}$	1. Tested hierarchy without any prior instruction. 2. Forgetting parameter not considered. 3. Unknown sequence of skill acquisition. 4. Small sample size 5. MSA restricted to independent linear scales. 6. Lacks sampling distribution and thus, generalizability. 7. Unable to estimate errors of measurement.
Airasian 1971	CONDITIONAL ITEM DIFFICULTY INDEX $P_{i+1 \mid i} = \frac{P_{i+1}}{P_i}$ where P_i = proportion who attain expected response pattern	1. Highly dependent on item validity. 2. Sensitive to sequencing of instruction. 3. Unable to estimate errors of measurement. 4. Forgetting/guessing parameters omitted.
Caple and Jones 1971	Phi-correlation coefficient	1. Estimations are highly dependent on test validity and reliability. 2. Existence of a hierarchy must yield high phi values; but high phi values do not necessarily establish a hierarchy.

Table 1 (continued)

Author	Statistical Techniques Employed	Weaknesses in the Model
Walbesser and Eisenberg 1972	DEPENDENCY TEST RATIOS: Consistency Ratio = $\frac{(11)}{(11)+(10)}$ Adequacy Ratio = $\frac{(11)}{(01)+(11)}$ Completeness Ratio = $\frac{(11)}{(11)+(00)}$ Necessity Ratio = $\frac{(00)}{(00)+(10)}$ Inverse Consistency Ratio = $\frac{(00)}{(00)+(01)}$	1. Errors of measurement suppress both C/R and N/R. 2. N/R particularly sensitive to the number of failures on the test. 3. Lacks generalizability. 4. Suggests invalid connections because of high ratios. 5. Indices measure how easy the test was for the examinees.
Bart, Airasian and Krus 1973	ORDERING-THEORETIC MODEL Percent of disconfirmatory response patterns (01) with pre-established tolerance levels.	1. Limited to pairwise analysis. 2. Depends upon pre-established tolerance levels for dealing with error (deterministic rather than probabilistic).
White and Clark 1973	PROBABILISTIC MODEL Estimates the probability that the examinee will be classified in the (01) cell, which is defined as the sum, over the four possible groups, of an examinee's probability of being in each group (P_0, P_I, P_{II}, P_B), multiplied by the conditional probability of members of the group answering the relevant number of items correctly: $P_{On} = P_0(1-\theta_b)^n \theta_d^n + P_I(1-\theta_a)^n\theta_d^n$ $+ P_{II}(1-\theta_b)^n \theta_c^n + P_B(1-\theta_a)^n \theta_c^n$	1. Valid connections may be rejected if sample size is too large. 2. Tests pairwise connections only.
Dayton and Macready 1976	MAXIMUM LIKLIHOOD MODEL A class of probabilistic models employing maximum liklihood estimates: $P_{(u)} = \sum_{j=1}^{q} P(u \mid v_j) \cdot a_j$, where $P(u \mid v_j) = \prod_{i=1}^{k} \alpha_i^{a_{ij}}(1-\alpha_i)^{b_{ij}} \beta_i^{c_{ij}} (1-\beta_i)^{d_{ij}}$ where α_i = guessing parameter and β_i = forgetting parameter	1. Limited to one item per skill.

Table 2

Comparison of Two Methodologies for Validating Hierarchies:

White & Clark vs. Dayton & Macready

Model	Advantages	Disadvantages
White & Clark	Uses multi-item measures for each objective	Valid connections may be rejected if sample size too large
		Tests pairwise connections only
		Probability estimates not maximum likelihood
Dayton & Macready:	Tests whole hierarchy	Limited to one item per objective
Models I and II	Maximum likelihood estimates	
	χ^2 goodness-of-fit	
	Allows for arbitrary hierarchies, both linear and branching	
	Allows for mastery/non-mastery decisions	
	Suggests hierarchy revisions as well as rejections	
Model III	Tests whole hierarchy	Limited to one item per objective
	Maximum likelihood estimates	
	χ^2 goodness-of-fit	
	Allows for arbitrary hierarchies, both linear and branching	
	Suggests hierarchy revisions as well as rejections	

Table 3

Thirteen Finally Selected Skills Utilized in the Study

Data Set[a]	Objective Code	Objective
I	22	Beginning consonant sounds
	24	Auditory discrimination: rhyming
(N=2319)	25	Ending consonant sounds
	27	Vowel sounds
	41	Beginning consonant digraphs
	43	Ending consonant digraphs
IV	45	Inflected/derived from root word
	48	Syllabication by vowel sound
(N=1118)	55	Prefixes, suffixes
	56	Root word + affix
V	55	Prefixes, suffixes
	56	Root word + affix
(N=1688)	63	Syllabication of nonsense word
	64	Suffixes and syntax (verbs)
	65	Suffixes and syntax (nouns)

[a]Data Sets II, III, and VI were eliminated because of the limited number of objectives in the cluster.

Table 4

Collapsed Responses to Hierarchy Specification Task For All Objective-Pairs Judged by the Experts (N=21)

Objective	Judgment			
A-B	A prior to B	B prior to A	Each facili-tates other	Order Irrelevant
24-22	7	4	3	7
22-25	17	0	1	3
41-22	1	16	3	1
22-43	17	0	1	3
27-22	2	9	1	9
24-25	12	2	1	6
24-41	9	1	3	8
24-43	10	2	2	7
24-27	13	2	1	5
25-41	9	1	3	8
43-25	1	11	5	3
25-27	5	0	4	12
43-41	2	9	7	3
41-27	2	0	2	17
27-43	3	2	4	12
45-48	1	5	1	14
45-55	4	8	7	2
45-56	7	4	8	2
48-55	3	2	0	15
48-56	2	2	1	16
55-56	12	4	5	0
55-56	12	4	5	0
63-55	4	2	4	11
55-64	12	4	5	0
65-55	5	11	5	0
56-63	2	5	1	13
64-56	4	9	5	3
56-65	8	2	8	3
63-64	4	3	2	12
63-65	3	3	0	15
64-65	1	6	7	7

Table 5

Percent of Agreement Between Judgment Data and Proposed Judgment Hierarchies

Data Set	Objective Code	Hierarchy a	b	c
I	43	56.2%	56.2%	56.2%
	41	57.1	57.1	57.1
	25	58.1	58.1	58.1
	24	48.6	38.1	48.6
	27	60.0	52.4	60.0
	22	62.9	60.0	62.9
	Average	57.2	53.7	57.2
IV	56	55.5		
	45	46.0		
	55	55.5		
	Average	52.3		
V	64	25.0		
	65	47.6		
	56	50.0		
	55	54.8		
	Average	44.4		

Table 6

Maximum Likelihood Estimates of Parameters and Their Standard Errors for All Data Sets at Each Criterion Level

	Data Set I				Data Set IV				Data Set V			
	n-1		n-2		n-1		n-2		n-1		n-2	
Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
θ_1[a]	0.016	0.41E-02	0.011	0.303E-02	0.999	0.372E-01	0.999	0.259E-01	0.965	0.358E-01	0.000	0.163E+17
θ_2	0.066	0.776E-02	0.029	0.509E-00	0.000	0.381E-01	0.000	0.268E-01	0.034	0.389E-01	0.000	0.212E+17
θ_3	0.009	0.588E-02	0.000	0.358E-02	0.000	0.263E-01	0.000	0.160E-01	0.000	0.996E-02	0.000	0.212E+17
θ_4	0.043	0.792E-02	0.029	0.582E-02	0.036	0.380E-01	0.000	0.132E-01	0.000	0.490E-02	0.000	0.212E+17
θ_5	0.245	0.148E-01	0.270	0.154E-01	--	--	--	--	0.000	0.252E-02	0.000	0.212E+17
θ_6	0.050	0.120E-01	0.000	0.112E-01	--	--	--	--	--	--	--	--
α[b]	0.257	0.163E-01	0.293	0.189E-01	0.503	0.113E-01	0.654	0.981E-02	0.757	0.586E-02	0.999	0.537E+08
β[c]	0.059	0.313E-02	0.029	0.206E-02	0.999	0.185E+00	0.999	0.186E+04	0.999	0.139E+00	0.999	0.126E+02

[a] θ_n is defined as the proportion at the nth true score pattern.

[b] α is the guessing probability.

[c] β is the forgetting probability.

Table 7

Observed and Predicted Frequencies, Chi-Square and Probability Estimates for True Score Patterns in Data Sets I, IV, and V

		n-1 Criterion				*n*-2 Criterion			
DATA SET	PATTERN	OBS N	PRED N	CHI-SQUARE	PROB	OBS N	PRED N	CHI-SQUARE	PROB
I	000000	10	8.29	.35	.0036	1	3.60	1.88	.0016
	100000	28	35.45	1.56	.0153	10	12.75	.59	.0055
	110000	28	20.75	2.53	.0089	5	6.12	.20	.0026
	111000	52	55.94	.27	.0241	33	30.76	.16	.0133
	111100	269	267.72	.00	.1154	298	288.80	.29	.1245
	111110	184	211.11	3.48	.0910	109	157.43	14.90	.0679
	111111	1036	974.61	3.86	.4203	1417	1340.22	4.39	.5779
IV	0000	34	48.06	4.11	.0430	11	15.92	1.52	.0142
	1000	131	69.08	55.48	.0618	58	30.17	25.65	.0270
	1100	200	69.87	242.35	.0625	151	57.16	154.01	.0511
	1110	26	70.66	28.23	.0632	43	108.29	39.37	.0969
	1111	131	71.46	49.58	.0639	306	205.16	49.55	.1835
V	00000	9	3.99	6.25	.0024	1	1687.98	1685.98	1.0000
	10000	4	4.29	.01	.0025	2	.00	2365.69	.0000
	11000	5	13.37	5.24	.0079	3	.00	*******	.0000
	11100	53	41.68	3.07	.0247	16	.00	*******	.0000
	11110	162	129.93	7.91	.0770	110	.00	*******	.0000
	11111	384	405.04	1.09	.2400	927	.00	*******	.0000

Table 8

Chi-Square Estimates for All Data Sets at Each Criterion Level

Statistic	Data Set I		Data Set IV		Data Set V	
	n-1 Criterion	*n*-2 Criterion	*n*-1 Criterion	*n*-2 Criterion	*n*-1 Criterion	*n*-2 Criterion
Likelihood ratio	0.352 E+03	0.185 E+03	0.161 E+04	0.120 E+04	0.167 E+04	0.539 E+05
Pearson χ^2	412.56	217.22	1899.9	1243.6	2982.3	----[a]
df	55	55	9	9	24	24

[a]Value of statistic exceeded the limits of the program.

Figure Captions

Figure 1. A 3x3 response matrix for the two-item case.

Figure 2. Proposed expert judgment hierarchies for Data Sets I, IV and V based on response data in Table 4.

Figure 3. Proposed Dayton & Macready hierarchies for Data Sets I, IV, and V based on discrete chi-square values of true score patterns found in Table 7.

Figure 4. Proposed White & Clark hierarchies for Data Sets I, IV, and V based on the observed frequencies for each objective-pair.

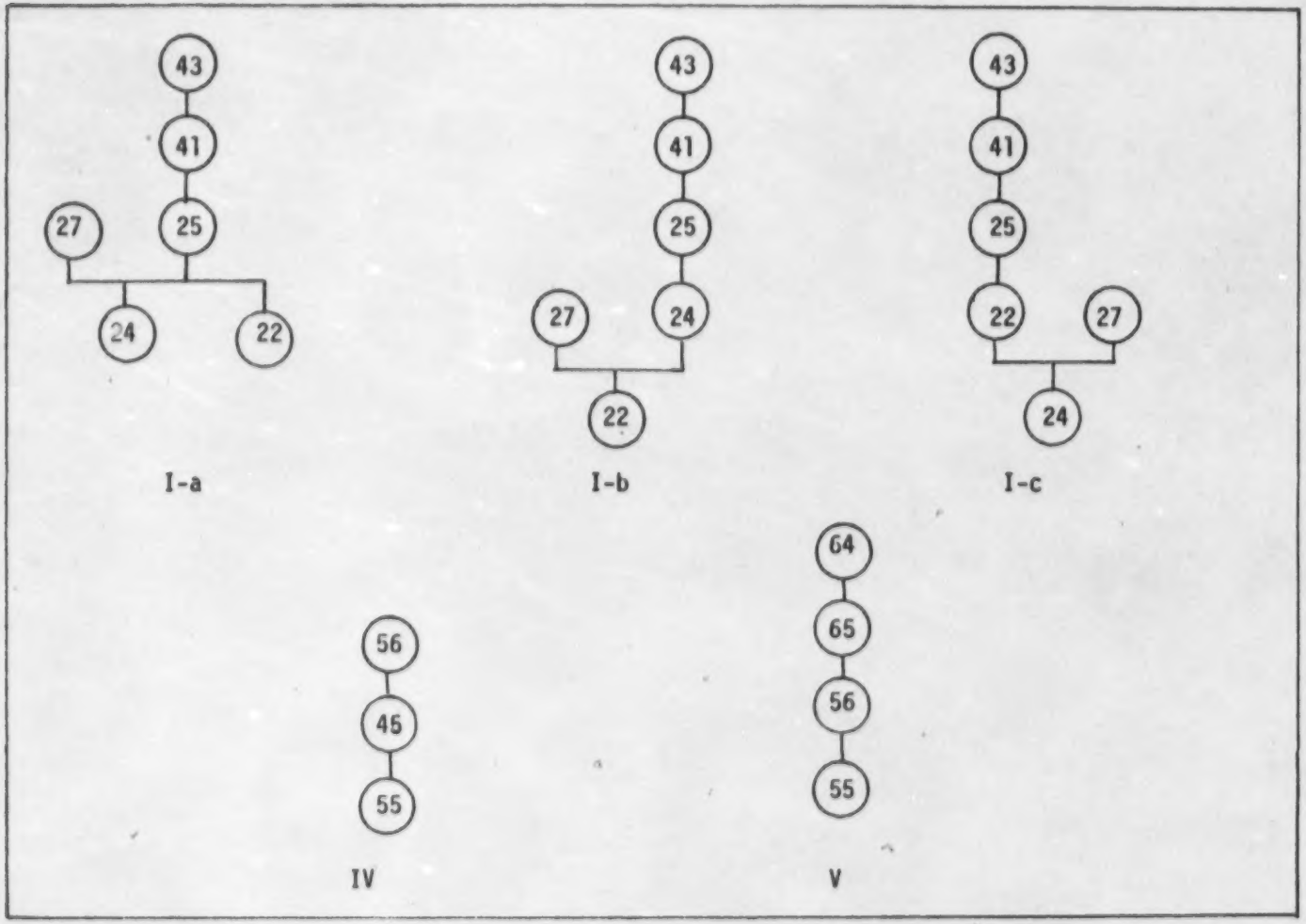
Figure 5. Proposed hierarchies for Dat Set I resulting from three methodological approaches.

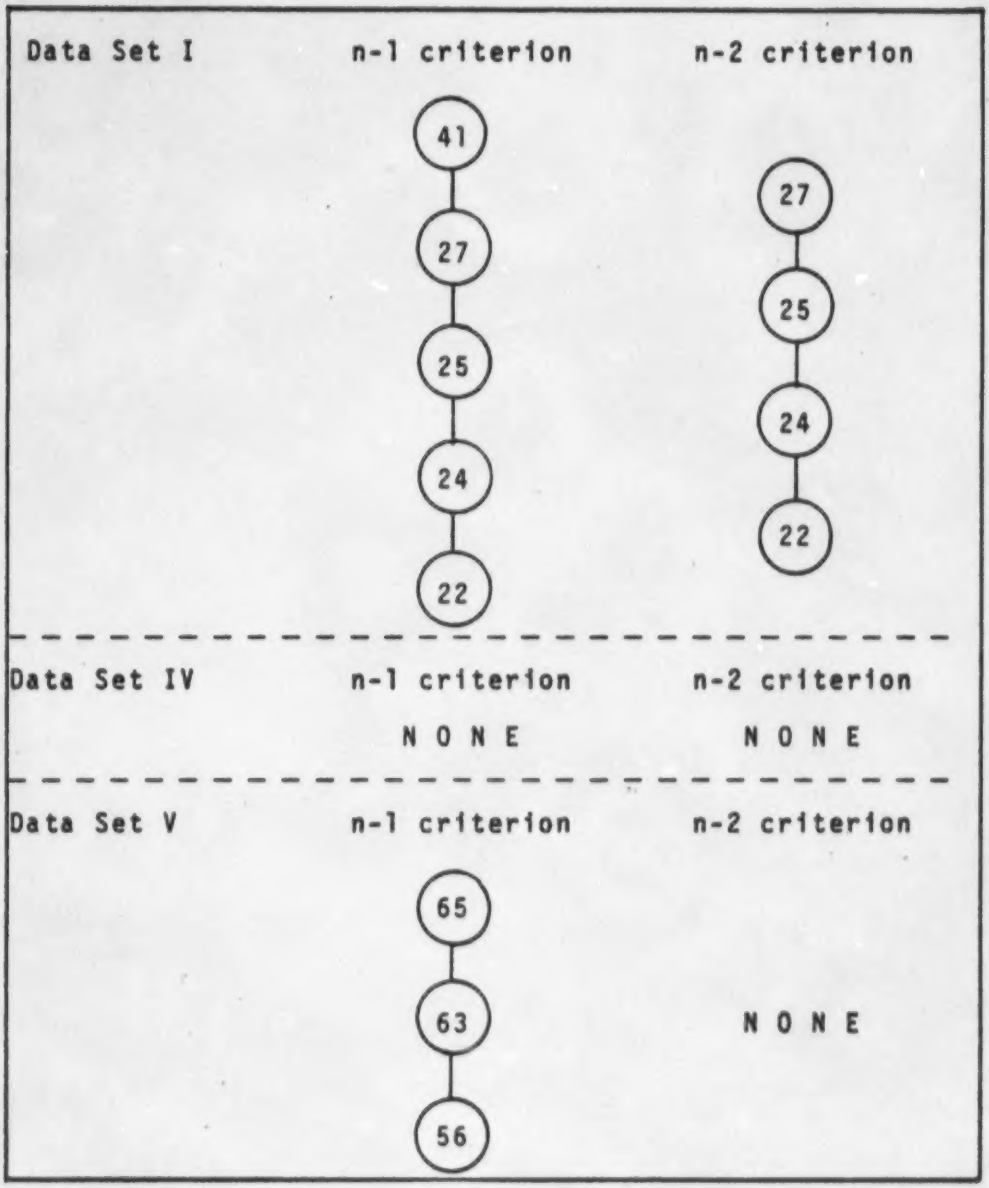
Figure 6. Proposed hierarchies for Data Set IV resulting from three methodological approaches.

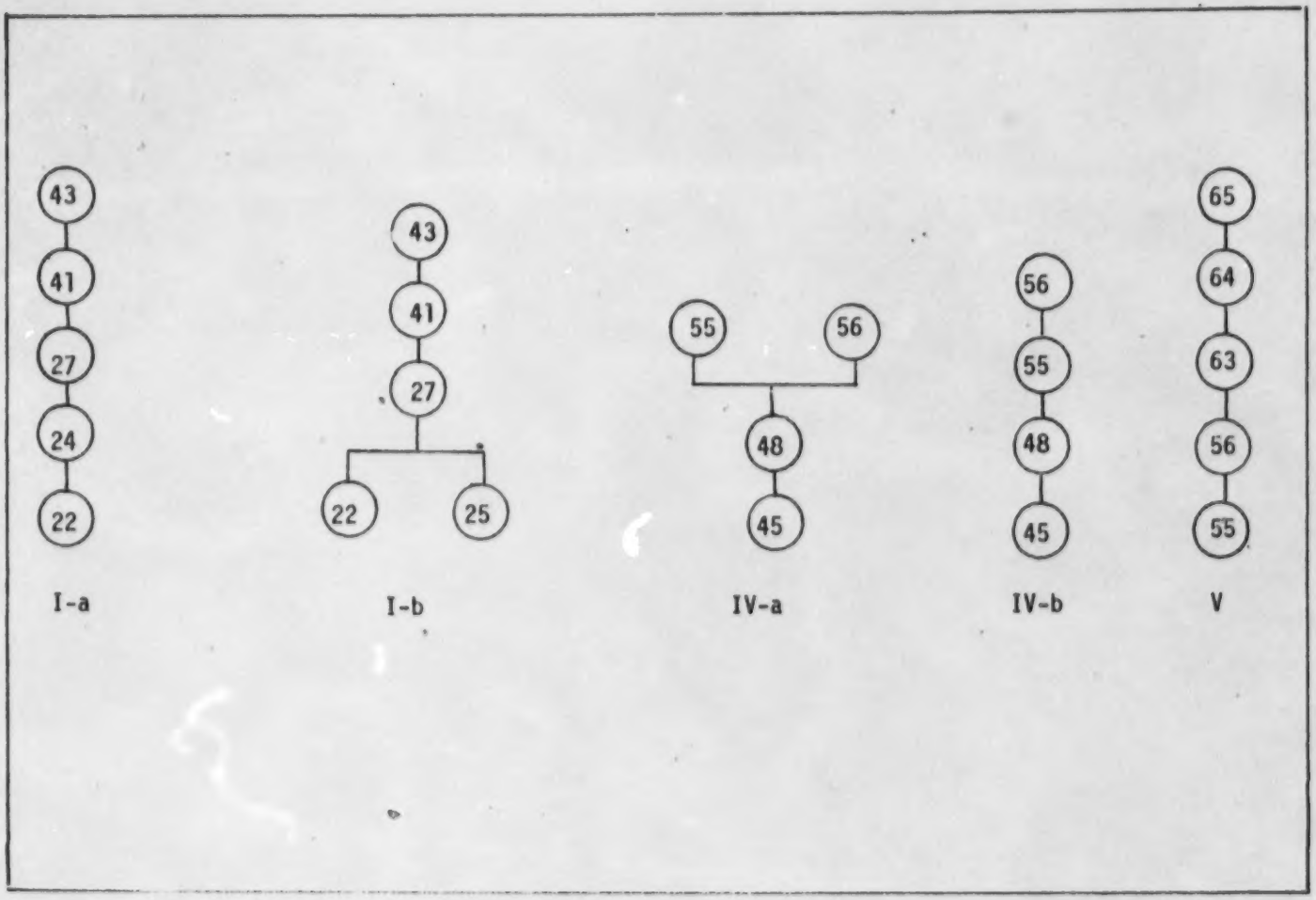
Figure 7. Proposed hierarchies for Data Set V resulting from three methodological approaches.

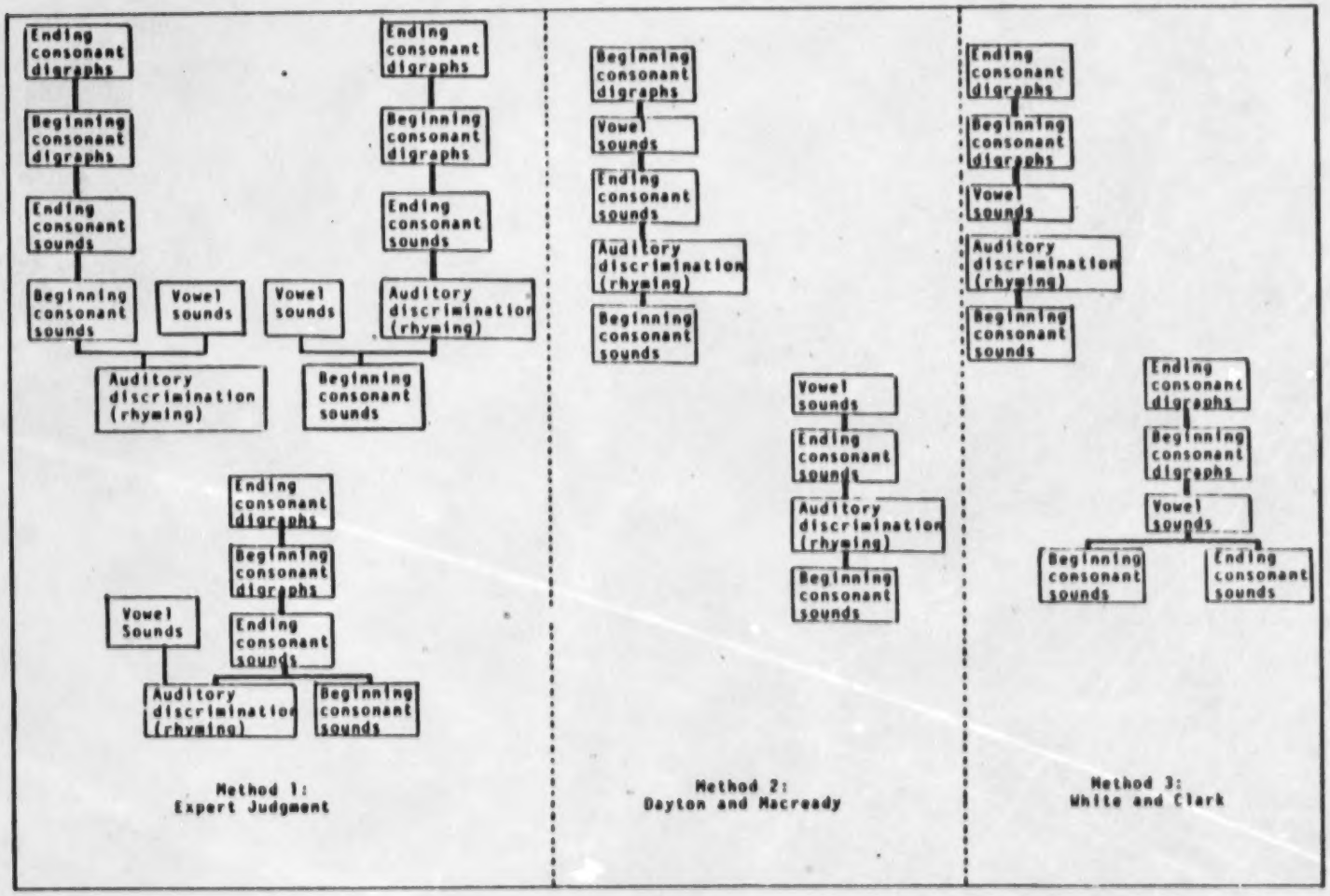
Skill II

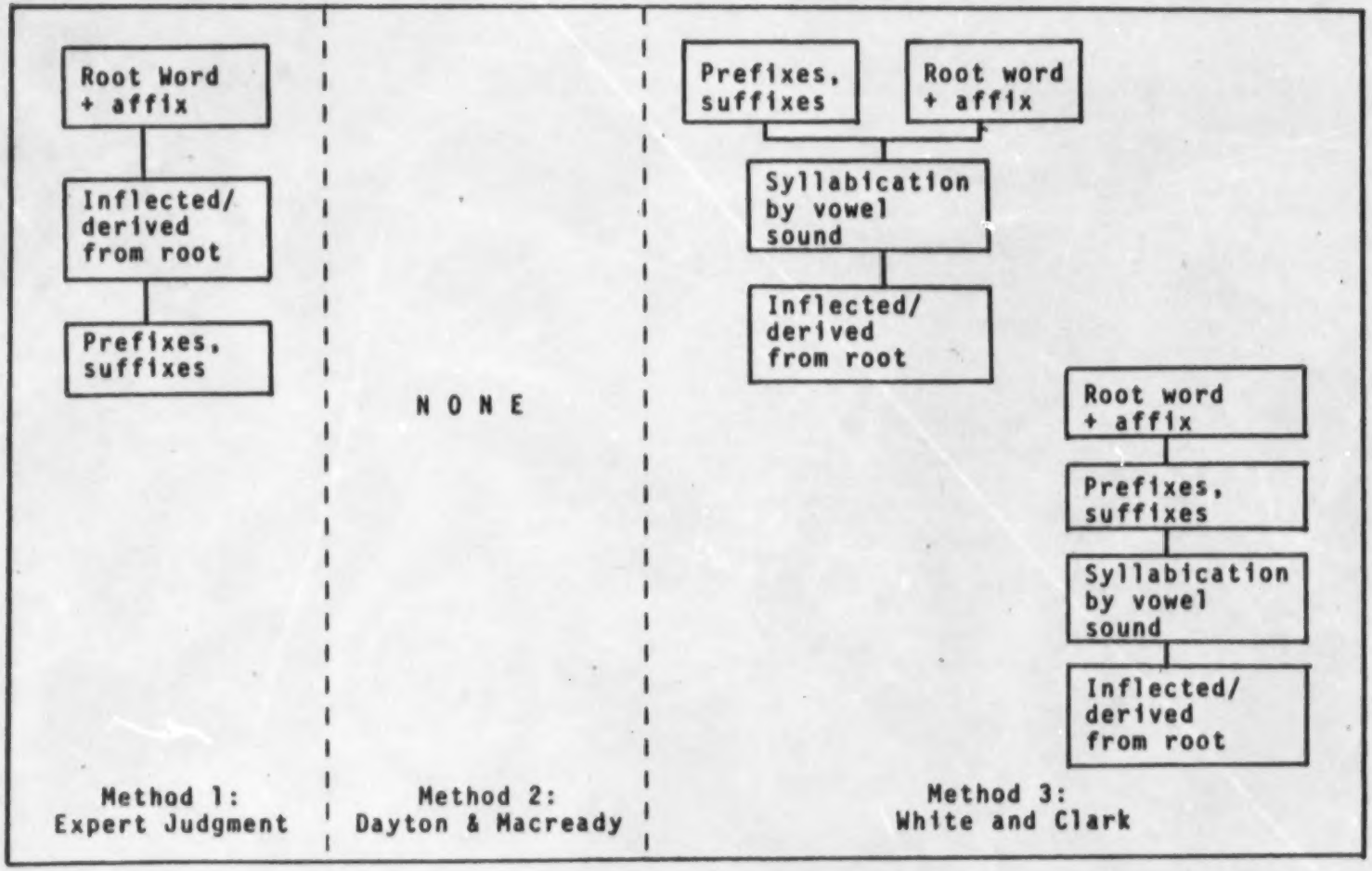
Skill I: Questions Correct	0	1	2
2	P_{20}	P_{21}	P_{22}
1	P_{10}	P_{11}	P_{12}
0	P_{00}	P_{01}	P_{02}

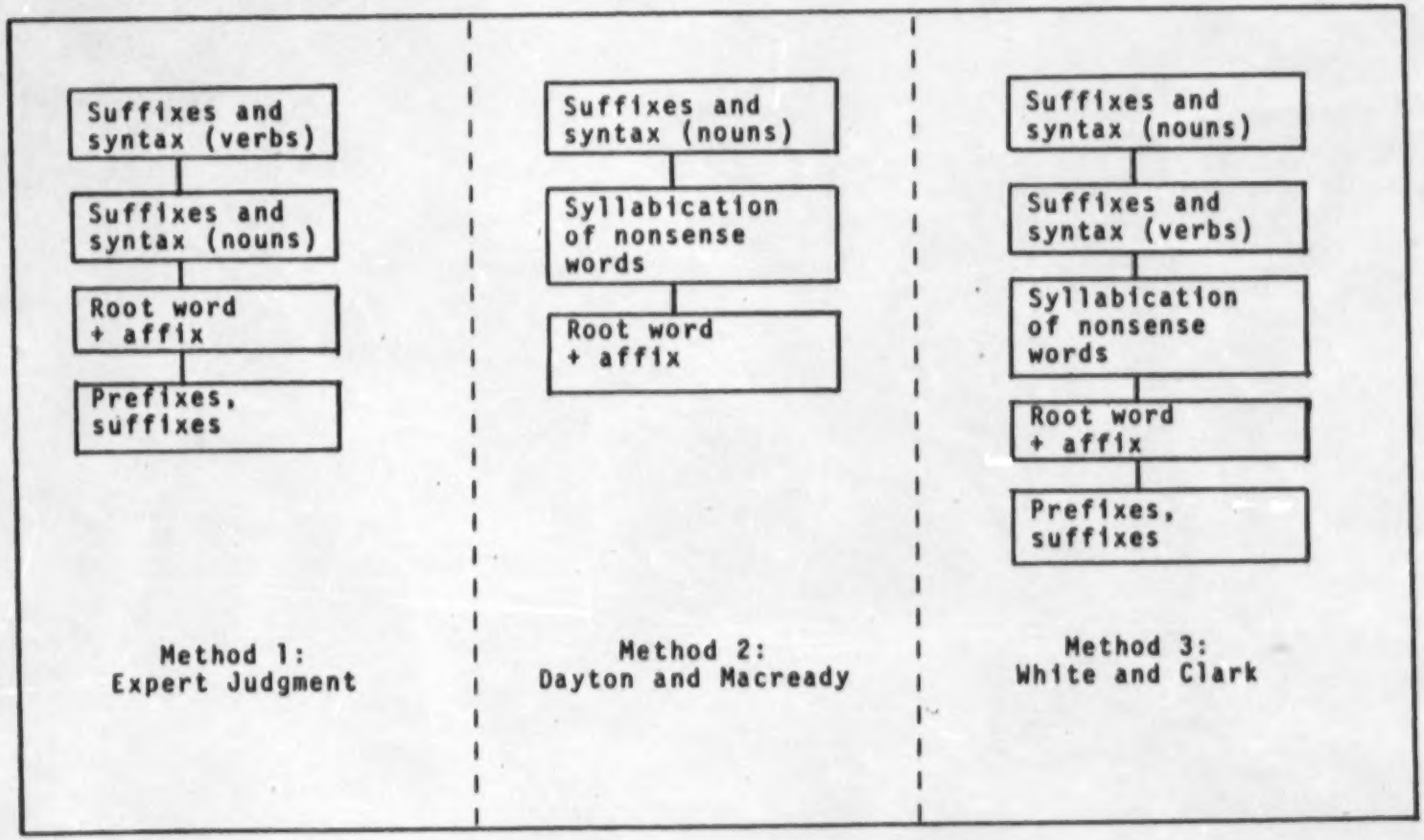

43
41
27
25
24
22
I-a
43
41
25
27
24
22
I-b
43
41
25
22
27
24
I-c
56
45
55
IV
64
65
56
55
V


Data Set I
n-1 criterion
n-2 criterion
41
27
25
24
22
27
25
24
22
Data Set IV
n-1 criterion
n-2 criterion
N O N E
N O N E
Data Set V
n-1 criterion
n-2 criterion
65
63
56
N O N E


43
41
27
24
22
I-a
43
41
27
22
25
I-b
55
56
48
45
IV-a
56
55
48
45
IV-b
65
64
63
56
55
V


Ending consonant digraphs
Beginning consonant digraphs
Ending consonant sounds
Beginning consonant sounds
Vowel sounds
Auditory discrimination (rhyming)
Method 1:
Expert Judgment
Method 2:
Dayton and Macready
Method 3:
White and Clark


Root Word + affix
Inflected/ derived from root
Prefixes, suffixes
Method 1: Expert Judgment
N O N E
Method 2: Dayton & Macready
Prefixes, suffixes
Root word + affix
Syllabication by vowel sound
Inflected/ derived from root
Root word + affix
Prefixes, suffixes
Syllabication by vowel sound
Inflected/ derived from root
Method 3: White and Clark


Suffixes and syntax (verbs)
Suffixes and syntax (nouns)
Root word + affix
Prefixes, suffixes
Method 1: Expert Judgment
Suffixes and syntax (nouns)
Syllabication of nonsense words
Root word + affix
Method 2: Dayton and Macready
Suffixes and syntax (nouns)
Suffixes and syntax (verbs)
Syllabication of nonsense words
Root word + affix
Prefixes, suffixes
Method 3: White and Clark